International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Important Feature Recognition for Credit Card Recommendation System using Predictive Modelling

Niti Desai[1], Neel Kothari[2*], Pratik Kanani[3], Bhoomi Shah[4], Lakshmi Kurup[5], Dashrath Kale[6], Nikita Raichada[7]

Submitted: 20/11/2023 **Revised**: 08/01/2024 **Accepted**: 22/01/2024

Abstract: The most popular electronic payment method is a credit card, which is more susceptible to theft due to the rising number of daily electronic transactions. Credit card frauds have cause credit card companies incur huge financial losses. To develop an effective credit scoring model is very imminent to prevent this fraud. In order to support the financial decisions made by banks and other financial organizations, researchers have created sophisticated credit scoring models using statistical and machine-learning techniques. Thus, the main aim of this paper is to help the bank management to develop models and predict consumer behavior on the basis of real time demographic data for credit card issuance. The research also exhibits how to treat data imbalance problem using Synthetic Minority Oversampling Technique (SMOTE) after applying various statistical tests. Different prediction models like Linear Regression, Decision Tree, XGBoost, AdaBoost, Random Forest etc. are also explored and applied on data to pick the best optimized one giving 92.03% accuracy and 97.32% Area Under the ROC Curve (AUC). The relevant parameters which are actually responsible for the identification of credit card fraud are highlighted by applying Weight of Evidence (WoE) and Inflation Variance (IV) techniques to all independent variables, which are able to find parameters having strong predicting power. The findings of such an experimental study can be really useful to bank managers to issue credit cards to customers.

Keywords: *SMOTE, Data imbalance, credit card, predictive modelling, recommendation system*

1. Introduction

The prevalence and use of credit cards have increased significantly since the early 2000's. Consumers typically, act logically to maximise their personal utility in the increasingly competitive credit card industry. However, it continues to be demonstrated that people abuse their credit cards, commit glaring errors and exhibit a variety of behavioural biases. Customers meeting certain profitable standards should receive distinct treatment depending on their repayment and purchase behaviour and can be awarded additional incentives [1]. Banks need to distinguish between customers who are likely to pay back and those who may not, leading to defaults. One of the ways we can achieve consumer prediction is by creating a system which takes into consideration various features and parameters like income, profession, education, etc and on the basis of these parameter's a decision can be made regarding the issuance of a credit card to a customer. Credit scoring models play a pivotal role in mitigating credit card fraud and curbing financial losses. They enable early fraud detection, assess applicants' creditworthiness, and monitor transaction patterns. Real-world impacts include reduced false positives, prevention of large-scale frauds, and minimized chargebacks, ultimately safeguarding credit card companies from substantial financial losses. The paper's credit card recommendation system likely employs a combination of data analysis, machine learning, and customer profiling. It assesses users' financial profiles, spending habits, and preferences to match them with suitable credit card options. By leveraging these insights, the system offers tailored recommendations that align with users' needs and financial circumstances, facilitating informed decision-making and increasing the likelihood of successful credit card applications.

David J. et al. [2] introduces credit scoring as a method of determining whether an applicant will default on their debt payments or not. Estimating the revenue and profits over a consumer's lifecycle and performing profit scoring; aiding in the determination of the loan's terms; adjusting as much to shifting economic circumstances are some of the advantages stressed of having complex predictive models that can identify when customers won't pay or settle debts in [3].Therefore, the analysis and prediction of whether issuing a credit card to a consumer would lead to defaulters or tracking and analysing whether the consumer will be able to pay the debt on time, would lead to reduced losses for the banks.

The Synthetic Minority Oversampling Technique (SMOTE) is a machine-learning approach aimed at mitigating data imbalance issues, commonly encountered in credit scoring models and other domains. It works by generating synthetic instances of the minority

[1] *Mukesh Patel School of Technology Management and Engineering, NMIMS University, Mumbai, India*
ORCID ID : 0000-0001-7906-2236
[2] *Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*
ORCID ID : 0000-0001-8336-4800
[3] *Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*
ORCID ID : 0000-0002-6848-2507
[4] *The Maharaja Sayajirao University of Baroda, Vadodara, India*
ORCID ID : 0000-0001-6881-2919
[5] *Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*
ORCID ID : 0000-0003-1579-2242
[6] *Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*
ORCID ID : 0009-0009-7320-5940
[7] *Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*
ORCID ID : 0009-0005-0452-9408
** Corresponding Author Email: neelkothariak@gmail.com*

class in an imbalanced dataset. SMOTE identifies minority class data points and creates synthetic examples by interpolating between them and their nearest neighbours. By doing so, it effectively balances the dataset, addressing the bias that often arises when one class significantly outnumbers the other. In credit scoring, this is vital as it ensures the model doesn't favour the majority class (e.g., non-defaulters) excessively. SMOTE's application results in improved model training, reduced bias, enhanced sensitivity to the minority class, better risk assessment, and fewer false negatives. However, it requires careful parameter tuning to avoid overfitting and achieve optimal model performance. Selecting the most optimized prediction model, among options like Linear Regression, Decision Tree, XGBoost, AdaBoost, and Random Forest, is pivotal for credit card issuance predictions due to its direct impact on accuracy, risk assessment, and business decisions.

Machine learning is the most well-liked and commonly used technology because of its broad range and scope of applications. Therefore, in recent times both machine learning and statistical approaches are employed to determine the risks related to credit card applicants. A banking organisation may have benefit extraordinarily by improving the effectiveness of the credit scoring algorithm. However, extra and useless data are commonly found in credit scoring data which might hurt a model's performance. Therefore, in order to increase efficiency and decrease model complexity, important features need to be recognized and only these features should be chosen for the credit scoring data. Tripathi et.al [4] combined nine feature selection methods with sixteen classification methods and consequently illustrated the benefits of using these feature selection methods. Most recent research focuses on improving classifier accuracy for predicting credit card default, however the target in this paper is to illustrate the utilisation of predictive modelling for obtaining recommendations for issuing credit cards. For better predictions and increased accuracy, recognition and extraction of important features was achieved. This was particularly difficult, because of the presence of imbalance in the class. To tackle data imbalance problems, Synthetic Minority Over-Sampling Technique (SMOTE) technique was utilized and different prediction models like Linear Regression, Decision Tree, XG Boost, Ada Boost and Random Forest were applied on the dataset. By combining data pre-processing techniques and predictive models, credit card recommendation accuracy improves through synergies. Data preparation refines and structures input data, enhancing model performance. Accurate recommendations based on individual attributes and financial behaviour benefit users with tailored solutions and financial institutions with reduced default risks. Ethical considerations include privacy and bias; addressing these ensures responsible and fair decision-making. The primary objective of developing credit scoring models, which integrate statistical and machine learning techniques, is to accurately assess the creditworthiness of individuals and businesses for informed lending decisions. These models leverage data analysis and predictive capabilities to determine the likelihood of borrowers repaying their loans or credit card debts on time. Moreover, they play a crucial role in addressing the escalating risk of credit card theft associated with the surge in electronic transactions. By analyzing transaction data in real-time, monitoring for unusual patterns, and employing machine-learning for anomaly detection, these models enhance fraud detection and prevention efforts. They contribute to safeguarding financial institutions and cardholders by identifying potential fraud, reducing false positives, and proactively responding to emerging threats in the electronic payment landscape.

1.1. Types of data

Structured data: This type of data, as the name suggests is structured in its format. There are rows and columns that contain the data items.

i. Nominal/categorical: This type of data is stated by virtue of words. It does not have a numerical value. For example, A person maybe wearing a “red” or a “black” cap, where red/black are the data items having nominal data type

ii. Numerical: In this type of data, data items are represented by numerical values(number) and the difference is of importance. For example, 5000 is more than 1000.

iii.Ordinal: This type of data has a specific order. However consecutive data points that characterize the attribute that is being considered, do not necessarily have scaled relations. For example. Arjun’s player rating is 5 and Rohan’s is 1 does not mean Arjun is 5 times better than Rohan.

iv.Time-series: This type of data is characterized over a period of time. For example, Shoe sales between the years 2000-2004.

1.2. Types of Statistical Analysis

The statistical tests mentioned below explores the relationship of two variables as well as the depth of this relationship. Following are the different ways of analysis:

i. Univariant: When you conduct a univariate analysis, there is just one variable in the data you are examining. The most common kind of univariate analysis involves examining a variable's range, central tendency (median, mean, and mode), maximum and lowest values, and standard deviation.

ii. Bivariant: In a bivariate study, two variables are compared to look at their correlations. These elements could be interdependent or autonomous of one another. In bivariate analysis, there is always a Y value for each X value.

iii.Multivariant: In a multivariate analysis, the link between more than two variables is examined. Multivariate analysis is the statistical examination of information gathered on many dependent variables. The majority of multivariate analysis approaches are expansions of univariate and bivariate analysis. Other multivariate methods were developed specifically to address multivariate issues.

1.3. Metrics to suggest importance of an evidence

On bases of Significant levels and P-values it is decided which relations to keep. They are metrics to determine importance of an evidence.

i. Significance Levels: A statistical hypothesis test's "critical probability" measures the likelihood that an inference proving a discrepancy between a measured value and a given statistical expectation is correct.

ii. P-values: The observed or calculated significance level, or probability number. To evaluate hypotheses, p-values and significance levels are compared. A theory is supported more strongly by higher p-values.

The scope and content of paper follows a sequential method. First comes the Data collection and preparation part. This is followed by the Variable correlation study using statistical analysis. The model is then tested, after removing relations having P-Value less

than significant level. This is followed by applying modern dataset balancing techniques. We then train and test the selected model using various methods. Finally, this is concluded with best optimum method which is useful to to recommend customers whom to issue credit card which leads minimum risk

The following is a breakdown of the paper's structure: Literature review combined with research gap are illustrated in Section 2. The Methodology is showcased in Section 3. Section 4 contains the results and Section 5 depicts the conclusion.

2. Literature Review

Multiple systems have been made on credit card transactions, few of them are to recommend credit cards to individuals, other few are to predict defaulters in credit cards and the leftover majority systems exist to detect fraud in the credit card transactions.

Credit card data exists in the numerical form with multiple attributes like gender, income, credit limit, defaulter status, etc. There exist multiple ways to use, interpret and analyze this data. Few known computation techniques are Logistic Regression, Decision Tree, XGBoost, AdaBoost, and Random Forest. These techniques are responsible to predict/forecast the future behavior of the system/user. However, to have successful predictions, these techniques should be fed with apt input data. This data should have no missing values and less/no outliers as they influence the results and make the system biased. There are multiple ways to filter the input data to make it more suitable for the model. In this paper, few of the statistical methods are shown to improvise the input data and an overview of other techniques are given.

Abhimanyu et al. in [5] have proposed a model using a deep-learning approach to predict number of frauds in credit card transactions. They have also proposed one technique to tune the parameters and later they have applied this technique to deep learning models. This paper aimed to increase the accuracy of fraud prediction in credit card transactions, so that the financial institutions will have less loses.

Different parameters like merchants' acceptance to use credit cards and user using the credit card have a great influence on the e-business. Researcher, Ying Loke, in [6] proposes a determinant based approach to show that the merchant acceptance to credit card payments in the business depends on the total value of sale, type of business and merchants' personal background. So, allowing credit card transactions in the business are very vital to boost the business performances.

Quality of Recommendations are very important in the recommendation systems. Nowadays, the number of recommendation parameters have increased and each user has a personal set of recommendation parameter values. The development of such a large-scale recommendation system, considering a plethora of recommendation parameter values, to suggest product and services and that even change every second for a large community is not easy. S. Thakur et. al in [7] propose a Bayesian Belief based approach to achieve scalable recommendation systems.

Multiple attributes signify different levels of significance in a prediction output. Whenever the system is finding the output, it uses different parameters to have the best possible predictions. Few parameters are quite significant in the output equations while others have less/no significance. This also depends on the model of prediction. So, before training the model, one should find the significance of the attributes and make sure that the attributes having the more significance are only considered while training the models in order to achieve higher accuracy, scalability and faster processing times. Significance of an attribute and different methods to deal with it are explained and discussed in [8].

Recommendation systems are attribute based. And these attributes are region based. So indirectly recommendation systems vary based on person's region, religion, behavior, economy, style of living etc. One of such research is done by Hanudin Amin [9], which shows patronage factors of Malaysian customers for Islamic credit cards. In this paper important attributes of a user have been found and their significance noted.

Human behavior such as confidence, financial spending styles and technology knowledge also plays an important role in the credit card success. One important research [10] based on the heuristics of all these three parameters shows that around 25% of the credit card behavior success lies in the confidence factor which is intangible. Every system has its own advantages and disadvantages. Using credit cards have multiple advantages but it also increases financial stress and results in unsecured debts. Randy Hodson et. al. in [11] study the advantages and disadvantages of using credit card system. Income amount and economy class have influences on the financial stress. This stress can be for short duration or for the long durations. These factors play an important role to make decisions whether the person should be allowed to use the credit card or not. Just issuing a credit card will not ensure its success and fulfilment of its purpose. There are different risks involved in it. The most common risk is number of frauds happening with credit card users.

There are multiple ways out of which one of the way is to use the network graphs [12] which tracks user behavior and location patterns and if any strange behavior occurs then it will be deleted by the graph using a network analysis model and the transaction suspected for the fraud will be stopped.

2.1. Research gap

All above discussed techniques perform pattern analysis in credit card scenarios and deal with fraud detection and attribute significance. Some also have confidence as one of the major factors using the credit card system and it also mentions the financial stress and debts incurred to users due to credit cards. However, no such method is present which can be deployed in the banking system for credit card recommendation to a customer or to pre-process the input data, so that it can best fit the model to give the most accurate output. In this paper not only data pre-processing using different statistical techniques but also output prediction using different models has been implemented. If the correct recommendation of the credit card based on the most important attributes is given, that is whether a person should be given a credit card or not based, then the correct user can enjoy its benefits and the other user can be prevented from financial stress and debts.

3. Methodology

This research makes use of the predictive models for identifying the right customers for acquisition of credit cards. For better predictions, our dataset has been optimized by selecting the valuable features using various techniques. Fig. 1 represents a basic architectural pipeline structure of the work.

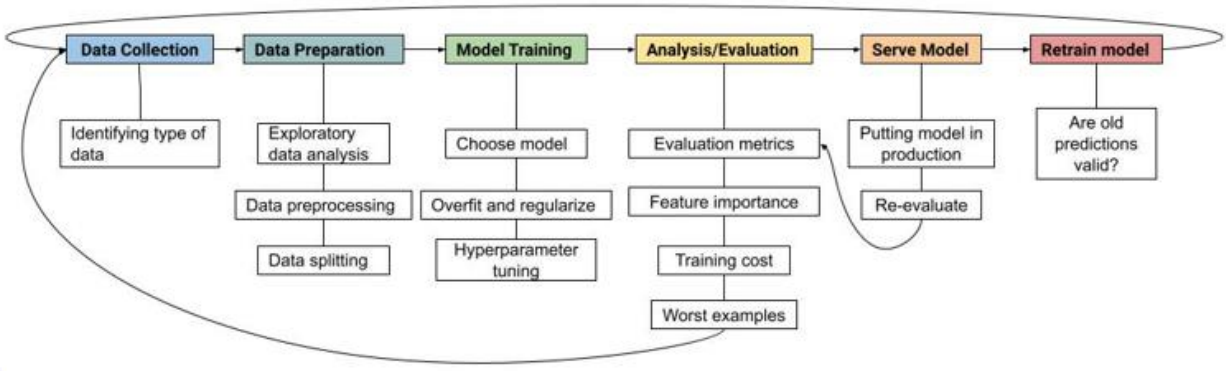

Fig. 1. Architectural pipeline structure

3.1. Data collection

The data collected is real-time data collected from ABC bank (name undisclosed). ABC is a leading international bank which offers a variety of banking services to customers. It gets thousands of credit card applications every year. Table 1 depicted below gives a brief description of the data used in this research. It reveals the data file type, number of files, source of data, etc.

Table 1. Data description

Properties	Status
Data Type	CSV files
No of Files	2
File Names	1." Credit Bureau data.csv" 2."Demographic data customer.csv
Source	Donated

Fig. 2 explains the various attributes of the Credit Bureau dataset used in the research where title refers to attribute and description is a short explanation of the attribute considered.

Fig. 3 explains the various attributes of the demographic data used in the research where title refers to attribute and description is a short explanation of the attribute considered.

Sr. No	Title	Description	Sr. No	Title	Description
1	Application ID	Unique ID of the customers	7	Education	Education of customers
2	Age	Age of customer	8	Profession	Profession of customers
3	Gender	Gender of customer	9	Type of residence	Type of residence of customers
4	Marital Status (at the time of application)	Marital status of customer (at the time of application)	10	No of months in current residence	No of months in current residence of customers
5	No of dependents	No. of children of customers	11	No of months in current company	No of months in current company of customers
6	Income	Income of customers	12	Performance Tag	performance (" 1 represents "Default")

Fig. 2. Attributes in Credit Bureau data

3.2. Data preparation and analysis

The first stage in any project is to gather the data in accordance with business requirements. The following stage involves pre-processing and cleaning the data, which includes removing values, managing imbalanced datasets, removing outliers, and converting categorical variables to numerical values, among other things. Fig. 4 shows the overview of the entire data exploration, analysis and visualization part of this research.

The entire data preparation process in broken down in a number of steps as mentioned below:

i. Step1: Merging of two datasets, where the common key is Application_ID. After merging two different data files, the merged dataset has 71301 rows, 30 attributes with 17 Integers, 8 float and 5 data objects. This is highlighted in green box in Fig. 4.

ii. Step 2: Exploratory data analysis (EDA), has been performed :

1. To identify the feature variables (input) and the target variable (output)
2. For Missing value treatment: To remove them or fill them feature imputation.
3. For Outliers treatment

After applying missing value detection (#Null/NA), 2596 rows are found affected, which is 3.65% of overall dataset.

Sr. No	Title	Description	Sr. No	Title	Description
1	Application ID	Customer application ID	11	No of PL trades opened in last 6 months	No of PL trades in last 6 months of customer
2	No of times 90 DPD or worse in last 6 months	Number of times customer has not paid dues since 90days in last 6 months	12	No of PL trades opened in last 12 months	No of PL trades in last 12 months of customer
3	No of times 60 DPD or worse in last 6 months	Number of times customer has not paid dues since 60 days last 6 months	13	No of Inquiries in last 6 months (excluding home & auto loans)	Number of times the customers has inquired in last 6 months
4	No of times 30 DPD or worse in last 6 months	Number of times customer has not paid dues since 30 days in last 6 months	14	No of Inquiries in last 12 months (excluding home & auto loans)	Number of times the customers has inquired in last 12 months
5	No of times 90 DPD or worse in last 12 months	Number of times customer has not paid dues since 90 days in last 12 months	15	Presence of open home loan	Does the customer have home loan (1 represents "Yes")
6	No of times 60 DPD or worse in last 12 months	Number of times customer has not paid dues since 60 days in last 12 months	16	Outstanding Balance	Outstanding balance of customer
7	No of times 30 DPD or worse in last 12 months	Number of times customer has not paid dues since 30 days in last 12 months	17	Total No of Trades	Number of times the customer has done total trades.
8	Avgas CC Utilization in last 12 months	Average utilization of credit card by customer	18	Presence of open auto loan	Does the customer have auto loan (1 represents "Yes")
9	No of trades opened in last 6 months	Number of times the customer has done the trades in last 6 months	19	Performance Tag	Status of customer performance (" 1 represents "Default")
10	No of trades opened in last 12 months	Number of times the customer has done the trades in last 12 months			

Fig. 3. Attributes in demographic data

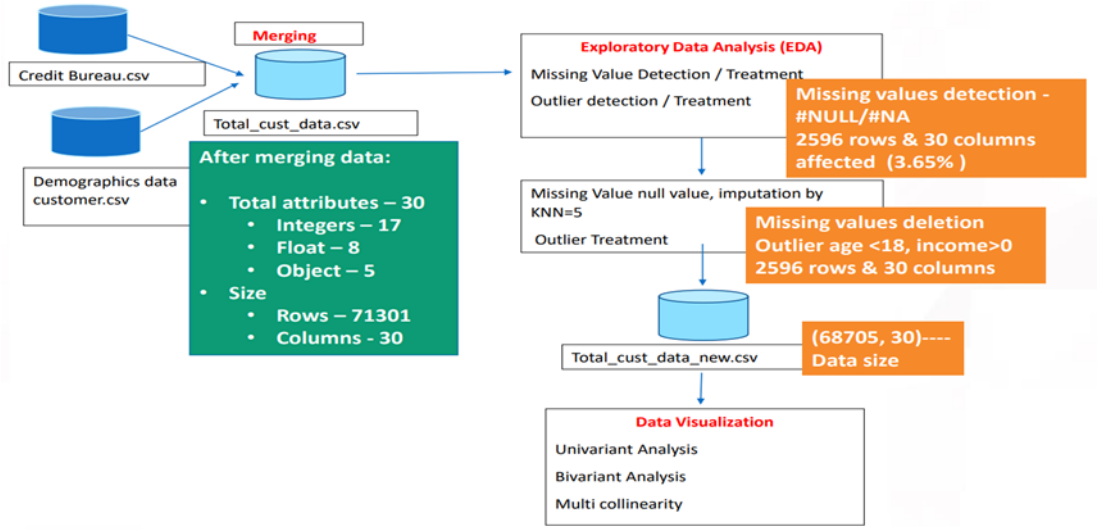

Fig. 4. Exploratory Data Analysis and visualisation

dropping missing values, within acceptable normal standard of data processing practice reduces number of rows from 71301 to 68705.

iii.Step 3: Feature imputation: filling missing values with mean, a median of the column.

For attributes having discrete values such as marital status, gender, etc. the missing values have been filled with mode of the respective column and for attributes having continuous values such as salary the median of the respective column is used.

Statistical Parameters	Age	No_of_Dpnd nts	Income	Tenure_Resi dence	Tenure_Com pany	Out_Balance	Perfrm_Y
count	68705	68705	68705	68705	68705	68,705	68705
mean	45.02	2.86	27.39	34.63	34.24	12,66,421	0.04
std	9.93	1.39	15.46	36.83	20.35	12,86,666	0.20
min	-3	1	-0.5	6	3	0	0
25%	38	2	14	6	17	2,13,450	0
50%	45	3	27	10	34	7,77,982	0
75%	53	4	40	61	51	29,30,258	0
max	65	5	60	126	133	52,18,801	1

Fig. 5. Outlier Detection after treating missing values

The Fig. 5 depicts outliers, highlighted by the yellow color. Outliers and illogical/erroneous values are highlighted in summary statistics of numerical variables.

3.2.1. Finding based correlation

Based on the Univariant, Bivariant and Multivariant analysis we identified the following few correlations and inferences in data:

i. Customers having high outstanding balance default more compared to others.

ii. Defaulting is linearly correlated to age i.e., as age increases there is a good chance of a customer defaulting.

iii.Customer income inversely affects defaulting i.e. as income increases there are less defaults.

iv.Total trade does not produce any pattern however it is observed that customers trading between 5 to 10 leads to high level of defaulting.

v. Number of dependents does not have any effect on defaulting.

vi.Customer residence stay has inverse effect on defaulting i.e., less is the stay, more is the default.

vii. Males are found to default more than Females.

viii. Education has no effect on defaulting.

ix.Salaried customers are found to default more than others.

x. A customer who is married is more probable to default.

xi.Customers staying on rent are found to default more than the

Sr.no	Original_Col_Name	Renamed_Col_Name
1	Application.ID	App_id
2	No.of.times.90.DPD.or.worse.in.last.6.months	N_90_DPD_6mnth
3	No.of.times.60.DPD.or.worse.in.last.6.months	N_60_DPD_6mnth
4	No.of.times.30.DPD.or.worse.in.last.6.months	N_30_DPD_6mnth
5	No.of.times.90.DPD.or.worse.in.last.12.months	N_90_DPD_12mnth
6	No.of.times.60.DPD.or.worse.in.last.12.months	N_60_DPD_12mnth
7	No.of.times.30.DPD.or.worse.in.last.12.months	N_30_DPD_12mnth
8	Avgas.CC.Utilization.in.last.12.months	Avg_CC_Utlzn_12mnth
9	No.of.trades.opened.in.last.6.months	N_trds_opn_6mnth
10	No.of.trades.opened.in.last.12.months	N_trds_opn_12mnth
11	No.of.PL.trades.opened.in.last.6.months	N_trds_PLopn_6mnth
12	No.of.PL.trades.opened.in.last.12.months	N_trds_PLopn_12mnth
13	No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	N_Inq_6mnth_Excl_AL
14	No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	N_Inq_12mnth_Excl_AL
15	Presence.of.open.home.loan	Opn_HmLoan
16	Outstanding.Balance	Ostng_Bal
17	Total.No.of.Trades	Total_Trades
18	Presence.of.open.auto.loan	Opn_AutLoan
19	Performance.Tag.x	PrmTag_X
20	Age	Age
21	Gender	Gender
22	Marital.Status..at.the.time.of.application.	Marital_Stat
23	No.of.dependents	N_Depndts
24	Income	Income
25	Education	Education
26	Profession	Profession
27	Type.of.residence	Residence_Type
28	No.of.months.in.current.residence	N_mnth_residence
29	No.of.months.in.current.company	N_mnth_CurCompany
30	Performance.Tag.y	PrmTag_Y

Fig. 6. Original Column name to Renamed column name mapping

ones who do not stay on rent.

3.2.2. Hypothesis Testing

We perform Hypothesis testing in order to assess the strength of the evidence

i. Null Hypothesis:

H0: There is no association between the twelve variables (refer Fig. 7) of the customer to become a credit card holder.

ii. Alternate Hypothesis:

H1: There is an association between the twelve variables (refer Fig. 7) of the customer to become a credit card holder.

The Fig. 6 contains a mapping of the original column name to the renamed column name used in the research.

After applying imputation and evaluation using ANOVA and Chi-

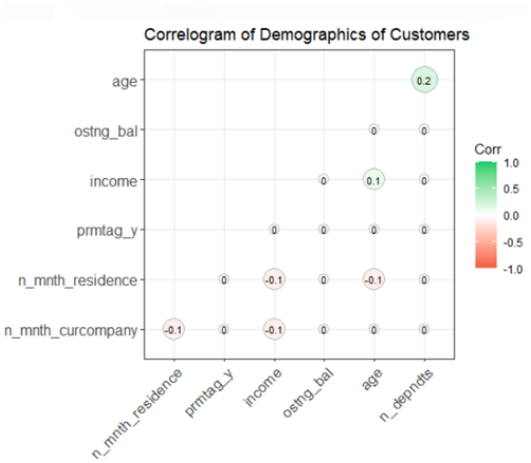

Fig. 7. Correlogram of Demographic of Customers

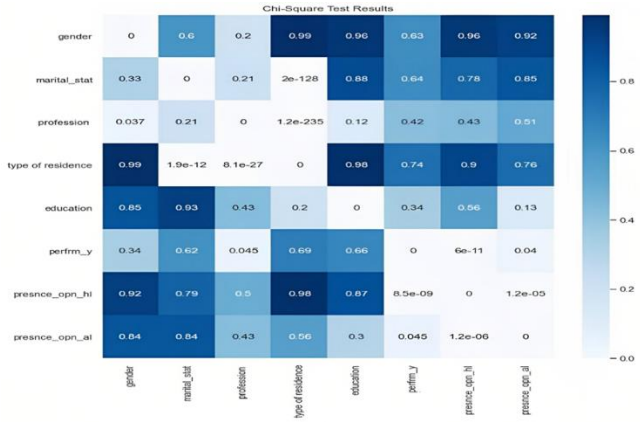

Fig. 8. Heat Map of Chi-Square test

Square test, it is observed that 20 on 27 attributes are more statistically significant having P value > 0.5 for Null Hypothesis H0. These results are shown in Table 2. We reject the null hypothesis when the p-value is less than or equal to your significance level. Conversely, when the p-value is greater than the significance level, you fail to reject the null hypothesis.

Table 2 shows the results of the ANOVA and Chi-Square tests.

Table 2. Result of ANOVA and Chi-Square Test

Feature	P_value	Null Hypothesis (Ho)	Test Type
Age	0.564	Accept	ANOVA
Out_balance	0.743	Accept	ANOVA
No of dependent	0.894	Accept	ANOVA
Type of residence	0.677	Accept	Chi-Square
education	0.638	Accept	Chi-Square
Marital status	0.63	Accept	Chi-Square
gender	0.34	Accept	Chi-Square

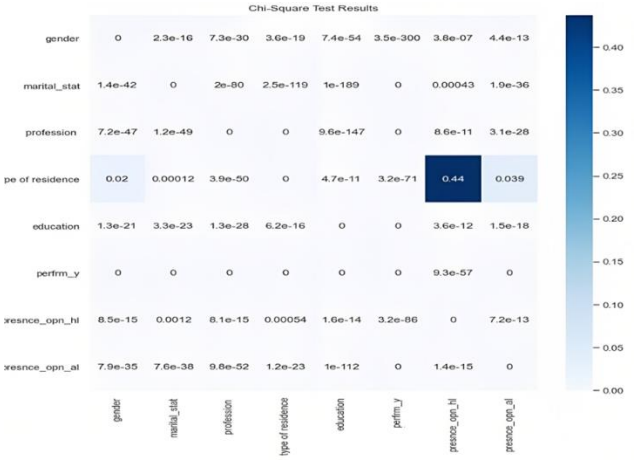

Fig. 9. Chi-Square test Results

Below Fig. 8, shows the heat map after applying Chi-Square test on various parameters.

3.2.3. Tables

3.3. Data Pre-processing

3.3.1. Limitations of Existing Dataset

After applying imputation techniques, we Cross-Validate data with 80%-20% train-test split. Model training accuracy of 95.88% has

been found. However, the dataset is highly imbalanced. So, accuracy of model for predicting issuance of credit card is very less i.e. 17.92%. This is significantly less as compared to model prediction accuracy. Therefore, it is necessary to balance the dataset. Multiple techniques like SMOTE, Random Over Sampling and Adaptive Synthetic Sampling can be applied. But to avoid duplicacy of data, SMOTE is preferred over other techniques. Table 3 shows the perfrm_y attribute values before and after applying SMOTE.

Table 3. Before and after applying Smote

0-> 66856

1-> 2946

Name: perfrm_y, dtype: int64

SMOTE	Class 0	Class 1
Before Applying	53453	2388
After Applying	53453	53453

SMOTE creates synthetic samples of minor classes tackling the data imbalance problem. In our case we had a dataset which consisted of 53453 instances for class 0 and 2388 instances for class 1. As the dataset was imbalanced, it consequently affected it's accuracy. The prediction for class 1 was 17.92% After applying SMOTE, prediction accuracy for 1 increased from 17.92% to 49.10%.

3.3.2. Statistical testing after applying SMOTE technique

After applying the SMOTE technique for balancing the data, ANOVA and Chi-Square tests are applied on Numerical and categorical data respectively for evaluation. Only 2 attributes are statistically significant shown in the Table 4 below. Rest of the attributes fail to accept H0 Hypothesis. The results of the Chi-square test are show in Fig. 9.

Table 4. Statistical testing after applying SMOTE

Feature	P_value	Null Hypothesis (0)	Test Type
90_dpd_6mnt	0.304	Accept	ANOVA
Out_balance	0.42	Accept	ANOVA

Weight of evidence (WoE): Weight of evidence (WoE) is a measure of how much the evidence supports or undermines a hypothesis. Weight of Evidence (WoE) and Inflation Variance (IV) techniques are used in credit card fraud detection to identify influential parameters. WoE measures the strength of a feature's association with fraud by analyzing the distribution of that feature for fraud and non-fraud cases. IV quantifies the information gain from each feature by assessing the differences in their distributions. Features with high WoE and IV values indicate strong predictive capabilities, helping pinpoint crucial attributes for fraud detection accurately. After applying Weight of Evidence (WoE) and IV to all independent variables, three attributes have been found that have strong predicting power. The variable name and IV score of these variables are mentioned in Table 5.

Table 5. Attribute having strong predicting power

Sr. no.	Variable	IV
1	tenure_residence	0.3342
2	avgas_cc_utilzn	0.4484
3	education	0.9919

Thus, as a whole, feature understanding is done by the following 6 ways:

i. Hypothesis Testing - Pearson's Chi-square test (Categorical 7 Features to Categorical Target variable) & ANOVA test (Continuous 20 Features to Categorical Target variable) This was performed before upsampling and after upsampling technique i.e. SMOTE.

ii. Correlation Matrix – Identified features bounded quantified relation with output variable and among themselves.

iii. Weight of evidence and Information Value (WOE & IV).

iv. Variance Inflation Factor - Checked collinearity.

v. Feature Selection Technique - Select K best class Top 10 best features observed.

vi. Feature Importance Index

4. Results

4.1. Simulation study and Training - Testing of Predictive Model

Initially, Logistic Regression (LR) is applied on the data. After applying Logistic Regression (LR1- base model) the accuracy found is 49% and the following things have been noticed:

i. 56 costumes defaulting but predicted as a performer.

ii. 10728 are actually performer predicted as a defaulter

Hence it was observed that feature selection and engineering are needed to improve the rate of prediction of the model.

Following methods are used in Feature selection:

i. WoE

ii. IV index

After applying feature selection accuracy is increased by 13.23% as compared to LR1(base model).

(refer Table 6.).

Table 6. LR model Accuracy

Logistic Regression	Change in Model	Accuracy
LR1	Base Model	49%
LR2	Removing weak prediction	62%
LR3	Removing weak + Medium predicator	62.23%

Table 7 shows the accuracy of the model without introducing feature importance.

Table 7. Accuracy of the model without feature importance

Name of Model	WoF removing weak predictor (Accuracy %)	Accuracy % after IV
Logistic Regression	62.23	72.95
Decision Tree	70	70
XGBoost	89	87.31
AdaBoost	81	80.9
Random Forest	44	42.3

Hyperparameter tuning is then applied by dropping a feature one at a time having high VIF (Variance Inflation Factor) and then the accuracy of XG model is tested. Table 8 shows the accuracy of the model after consecutively dropping attributes.

Table 8. Accuracy of XGBoost model after dropping high VIF attributes

Drop attribute	VIF Score	Accuracy
Trade in 12 months	107.5	86.06
Presence_opn_h	151.98	86.13
No_income_12mnts	15.75	86.25
No_income_6months	10	86.10
Type of Residence	15.83	85.87
age	19.50	85.21

Table 9 shows final accuracy after dropping features having high VIF. It gives 92% accuracy and an AUC of 97.32% which is 4.69% greater as compared to previous one.

Table 9. Final Model XGBoost

	precision	recall	f1-score	support
0	0.93	0.90	0.92	10715
1	0.91	0.94	0.92	10666
Accuracy			0.92	21381
Macro avg	0.92	0.92	0.92	21381
Weighted avg	0.92	0.92	0.92	21381

Optimised XGBoost classifier gives 92.03% accuracy and 97.32% AUC. Hyperparameter tuning is 80%.

5. Conclusion

Prediction of the customers who will likely be default in future and to identify them in advance can help the financial institutions. Most prevalent technique used to offer credit cards or not to customers is credit score rating. There are different techniques like Logistic Regression, Ada Boost, XG Boost, Random Forest etc. used for prediction. Here, these different techniques are compared with each other and it is found that XGBoost Ensemble method is ideal for regression and classification. Reduced biases and variance boost accuracy of XGBoost by 4.69%. The optimised XGBoost classifier gives 92.03% accuracy. Strong correlation of variables is also identified by statistical study on real time dataset by considering the significant level (>=0.05) of P-Value.

Following facts are also found based on real time dataset:

i. Number of dependents has no effect on delinquency whereas income factors negatively on delinquency

ii. Having education status like professional, masters, gender status as a male, marital status as married leads to delinquency more often as compared to their parallel respective counters.

iii. Analysis also showcased that less the tenure in the company more is delinquency. Rented customers will be defaulting more as compared to other residency types.

iv. Housing loans and DPD were found to have no casual relation or any correlation between defaulting and delinquency.

The findings in this research can be really useful to bank managers to issue credit cards to customers.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by all authors. The manuscript was written and revised by all authors on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Dyché, J. (2009) The CRM handbook: A business guide to customer relationship management. Boston, MA: Addison-Wesley..

[2] Hand, D.J. and Henley, W.E. (1997) 'Statistical Classification Methods in Consumer Credit scoring: A Review', Journal of the Royal Statistical Society Series A: Statistics in Society, 160(3), pp. 523–541. doi:10.1111/j.1467-985x.1997.00078.x.

[3] Baesens, B. et al. (2005) 'Neural network survival analysis for personal loan data', Journal of the Operational Research Society, 56(9), pp. 1089–1098. doi:10.1057/palgrave.jors.2601990.

[4] Tripathi, D. et al. (2021) 'Experimental Analysis of Machine Learning Methods for Credit Score Classification', Progress in Artificial Intelligence, 10(3), pp. 217–243. doi:10.1007/s13748-021-00238-2.

[5] Roy, A. et al. (2018) 'Deep learning detecting fraud in credit card transactions', 2018 Systems and Information Engineering Design Symposium (SIEDS) [Preprint]. doi:10.1109/sieds.2018.8374722.

[6] Loke, Y.J. (2007) 'Determinants of merchant participation in credit card payment schemes', Review of Network Economics, 6(4). doi:10.2202/1446-9022.1130.

[7] Thakur, S.S., Kundu, A. and Sing, J.K. (2011) 'A novel approach: Using bayesian belief networks in product recommendation', 2011 Second International Conference on Emerging Applications of Information Technology [Preprint]. doi:10.1109/eait.2011.21.

[8] Leong, L.-Y. et al. (2013) 'Predicting the determinants of the NFC-enabled mobile credit card acceptance: A Neural Networks approach', Expert Systems with Applications, 40(14), pp. 5604–5620. doi:10.1016/j.eswa.2013.04.018.

[9] Amin, H. (2012) 'Patronage factors of Malaysian local customers toward Islamic credit cards', Management Research Review, 35(6), pp. 512–530. doi:10.1108/01409171211238271.

[10] Shefrin, H. and Nicols, C.M. (2014) 'Credit card behavior, financial styles, and heuristics', Journal of Business Research, 67(8), pp. 1679–1687. doi:10.1016/j.jbusres.2014.02.014..

[11] Hodson, R., Dwyer, R.E. and Neilson, L.A. (2014) 'Credit card blues: The middle class and the hidden costs of Easy Credit', The Sociological Quarterly, 55(2), pp. 315–340. doi:10.1111/tsq.12059.

[12] Zanin, M. et al. (2018) 'Credit card fraud detection through PARENCLITIC network analysis', Complexity, 2018, pp. 1–9. doi:10.1155/2018/5764370.. W.-K. Chen, Linear Networks and Systems. Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.